Uncertainty Detection in Hungarian Texts

Veronika Vincze[1,2]

[1] MTA-SZTE, Research Group on Artificial Intelligence
[2] University of Szeged, Department of Informatics
vinczev@inf.u-szeged.hu

Distinguishing between factual (i.e. true or false) and uncertain propositions is essential both in linguistics and natural language processing applications. For instance, in information extraction (IE) many applications seek to extract factual information from text, and they should handle detected modified parts in a different manner. Due to this, uncertainty detection has received a considerable amount of attention in the last few years in the natural language processing community.

In this paper, we report on a Hungarian corpus – hUnCertainty – manually annotated for several types of linguistic uncertainty, which is – to the best of our knowledge – is the first one developed for Hungarian.

The hUnCertainty corpus contains paragraphs from the Hungarian Wikipedia. Hungarian equivalents of typical uncertainty cues in English were collected and paragraphs containing them were randomly sampled from the Hungarian Wikipedia dump. Besides, paragraphs which did not contain such words were also included in the corpus so as to avoid biased data.

The corpus is manually annotated for linguistic cues denoting several types of uncertainty. A sentence is epistemically uncertain if on the basis of our world knowledge we cannot decide at the moment whether it is true or false. As for hypothetical uncertainty, the truth value of the propositions cannot be determined either. This class contains conditionals and investigations, which is frequent in science papers where research questions are often stated in the form of this linguistic tool. Non-epistemic types of modality (such as doxastic modality – related to beliefs – or dynamic modality – related to e.g. necessities) also belong to this group.

Concerning discourse-level uncertainty, we annotated three classes. First, weasels are sourceless propositions or propositions with any underspecified argument that would be relevant or is not common knowledge in the situation. Second, hedges blur the exact meaning of some qualities or quantities. Third, peacocks express unprovable qualifications or exaggerations.

This corpus served as the training and test database for our CRF-based approach, which makes use of a rich feature set including orthographic, lexical, morphological, syntactic and semantic features as well. The results of our experiments show that uncertainty detection can be successfully carried out on Hungarian texts as well.